Response to comments

Paper #: essd-2020-182

Title: GLC_FCS30: GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery

Journal: Earth System Science Data

Reviewer #2

While I appreciate the authors' tremendous efforts in this global-scale mapping project, I have several major concerns.

Great thanks for the comment. The manuscript has been improved according to your and another reviewer's comments.

From the remote sensing perspective, the novelty of this project is low. Almost all the methods have been developed and used somewhere in the previous land-cover mapping projects.

Great thanks for the comment. As the global-scale mapping involves tremendous efforts and workloads, we split the project into three parts: 1) the work of "Fine Land-Cover Mapping in China Using Landsat Datacube and an Operational SPECLib-Based Approach" analyzed the accuracy and robustness of the automatic classification strategy; 2) the work of "Development of a global 30 m impervious surface map using multisource and multitemporal remote sensing datasets with the Google Earth Engine platform" used the multi-source and multi-temporal imagery to guarantee the high accuracy of impervious surfaces. 3) Based on our previous works (1-2), we combined the time-series Landsat imagery and GSPECLib to generate the GLC_FCS30-2015 global 30 m land-cover products. Therefore, we think the project is an **incremental innovation** because **the GLC_FCS30-2015 is the first global 30 m land-cover product based on an automatic classification strategy, and has significant advantages over mapping accuracy comparing with the current global 30 m products.**

Zhang, X., Liu, L., Chen, X., Xie, S., and Gao, Y.: Fine Land-Cover Mapping in China Using Landsat Datacube and an Operational SPECLib-Based Approach, Remote Sensing, 11, 1056, https://doi.org/10.3390/rs11091056, 2019.

Zhang, X., Liu, L., Wu, C., Chen, X., Gao, Y., Xie, S., and Zhang, B.: Development of a global 30 m impervious surface map using multisource and multitemporal remote sensing datasets with the Google Earth Engine platform, Earth Syst. Sci. Data, 12, 1625-1648, https://doi.org/10.5194/essd-12-1625-2020, 2020.

The classification system proposed in the study looks relatively simple. The study is not targeting the issue - "a fine land-cover system is still lacking" - as described at the end of the Introduction.

Great thanks for the comment. According to the reviewing in the introduction, the current global 30 m land-cover products mainly used the simple classification system (containing 10 major land-cover types), however, our GLC_FCS30-2015 products adopted the CCI_LC (Climate Change Initiative Global Land Cover) classification system containing 30 land-cover types, so it has significant advantages over land-cover diversity comparing with current global 30 m products (for example, only ten land-cover types in GlobeLand30). Based on the comment, the sentence has been deleted in the Introduction as:

"Overall, due to the difficulties in collecting sufficient accurate training data with a fine classification system and the computing requirements involved, producing a global 30-m land-cover classification with a fine classification system is a challenging and labor-intensive task."

However, the construction of the training database is a great effort that should be given more emphasis in the description of methods (e.g., adding a flowchart) and in the discussion (e.g., effects of sample outliers on mapping accuracies across land cover classes). See details below.

Great thanks for the comment. Based on the suggestion, the details of the deriving training samples have been added (the effects of sample outliers have been explained in the next comment)

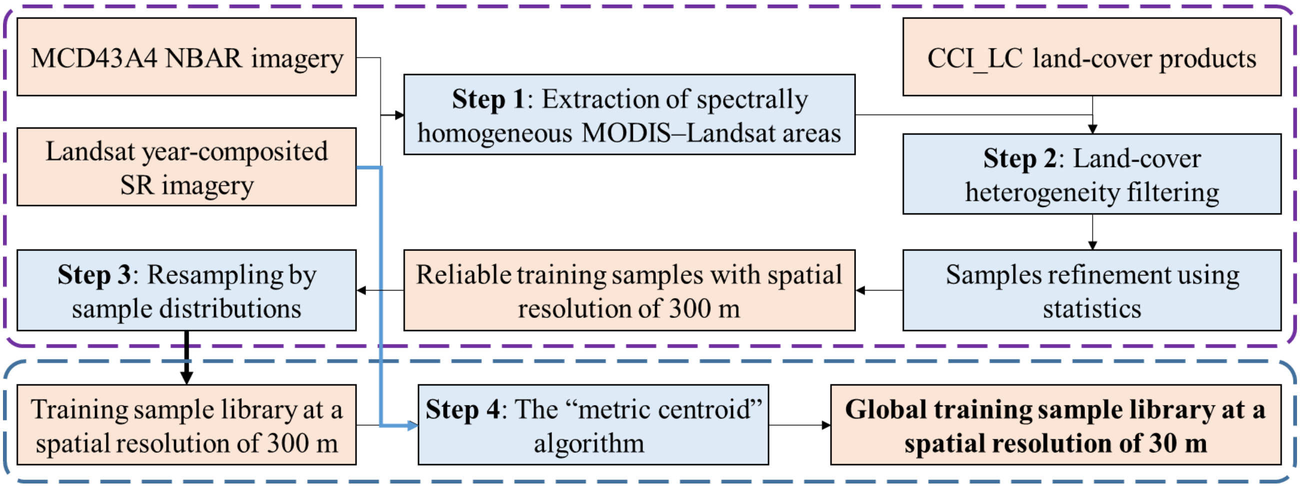
MCD43A4 NBAR imagery
Landsat year-composited SR imagery
Step 1: Extraction of spectrally homogeneous MODIS–Landsat areas
CCI_LC land-cover products
Step 2: Land-cover heterogeneity filtering
Step 3: Resampling by sample distributions
Reliable training samples with spatial resolution of 300 m
Samples refinement using statistics
Training sample library at a spatial resolution of 300 m
Step 4: The "metric centroid" algorithm
Global training sample library at a spatial resolution of 30 m

Figure 3. The flowchart of deriving training samples by using multi-source datasets.

Similar to our previous works (Zhang et al., 2019; Zhang et al., 2018), four key steps were adopted to guarantee the confidence of each training point, as illustrated in the Figure 3. As in Zhang et al. (2019), the spectrally homogeneous MODIS–Landsat areas were firstly identified based on the variance of a 3×3 local window using spectral thresholds of [0.03, 0.03, 0.03, 0.06, 0.03, and 0.03] for the six MODIS bands (blue, green, red, NIR, SWIR1, and SWIR2) in the both MCD43A4 NBAR products and Landsat SR imagery (Feng et al., 2012). It should be noted that the year-composited Landsat SR data were downloaded from GEE platform with the sinusoidal projection. As the MCD43A4 NBAR is corrected for view-angle effects and Landsat has a small view angle of ±7.5°, the view-angle difference between MCD43A4 and Landsat SR could be considered negligible.

Before the process of refinement and labeling, the CCI_LC land-cover products, which had geographical projections, were reprojected to the sinusoidal projection of MCD43A4. The spatial resolution of MCD43A4 was 1.67 times that of the CCI_LC land-cover product and the spectrally homogeneous MODIS–Landsat areas had been identified in the 3×3 local windows. Also, Defourny et al. (2018) and Yang et al. (2017b) found that the CCI_LC performed better over homogeneous areas; therefore, a larger local 5×5 window was applied to the CCI_LC land-cover product to refine and label each spectrally homogeneous MODIS-Landsat pixel. Specifically, the land-cover heterogeneity in the local 5×5 window was calculated as being the percentages of land-cover types occurring within the window (Jokar Arsanjani et al., 2016a). Aware of the possibility of reprojection and classification errors in the CCI_LC products, the land-cover heterogeneity threshold was empirically selected as approximately 0.95; in other words, if the maximum frequency of dominant land-cover types was less than 22 in the 5×5 window, the point was excluded from GSPECLib. After a spatial–spectral filter had been applied to MCD43A4 and a heterogeneity filter to the CCI_LC product, the points that had homogeneous spectra and land-cover types were retained. In addition, to further remove the abnormal points contaminating by classification error in the CCI_LC, the homogeneous points were refined based on their spectral statistics distribution, in which the normal samples would form the peak of the distribution whereas the influenced samples were on the long tail (Zhang et al., 2018). It should be noted that the geographical coordinates of each homogeneous point were selected as being the center of the local window in the CCI_LC product because this had a higher spatial resolution than that of MCD43A4.

Then, Zhu et al. (2016) and Jin et al. (2014) found that the distribution (proportional to area and equal allocation) and balance of training data had significant impact on classification results, and quantitatively demonstrated that the proportional approach usually achieve higher overall accuracy than the equal allocation distribution. In addition, Zhu et al. (2016) also suggested to extract a minimum of 600 training pixels and a maximum of 8000 training pixels per class for alleviating the problem of unbalancing training data. In this study, the proportional distribution and sample balancing parameters were used to resample these homogeneous points in each GSPECLib 158.85 km×158.85 km geographic grid cell.

Lastly, different from the previous spectrally based classification using MCD43A4 reflectance spectra (Zhang et al., 2019), in this study, we proposed to use the Landsat reflectance spectra , derived by combining the global training samples and time-series Landsat imagery, to produce the global 30 m land-cover mapping. However, as the spatial resolution difference between Landsat SR (30 m) and homogeneous training samples (300 m), therefore, the “metric centroid” algorithm proposed by Zhang and Roy (2017) was used to find the optimal and corresponding training points at a resolution of 30 m. Specifically, as each homogeneous point corresponded to an area equivalent to 10×10 Landsat pixels, the normalized distances (Eq. (2)) between each Landsat pixel and the mean of all 10×10 pixel areas were calculated. The optimal and corresponding training points at 30 m were selected as the ones having the minimum normalized distance,

$$D_i = \left(\rho_i - \frac{1}{n}\sum_{j=1}^{n}\rho_j\right)^2, i = 1,2,\dots,n \tag{2}$$

where ρ_i is a vector representing the annually composited Landsat SR for 2015 and n is the number of Landsat pixels within a 10×10 local window (defined as 100). If several 30-m pixels had the same minimum D_i value then one pixel was selected at random.

I also feel there is a lack of in-depth discussion. For a large-scale project, data uncertainties, model calibration, and land-cover heterogeneity could have a significant effect on mapping accuracy. But the current form of discussion is superficial and needs to add a comprehensive evaluation of the developed database.

Great thanks for the comment. Based on the suggestion and subsequent comments, the Discussion has been totally strengthened as:

1) As for the analysis of training data uncertainties

To demonstrate the importance of sample sizes, 200,000 points, approximately 1% of total training samples, were randomly selected to quantitatively analyse the relationship between overall accuracy and the corresponding sample size. Specifically, we used the 10-fold cross-validation method to split these points into training and validation samples, and then gradually increase the size of training samples with the step of 2% and repeat the process for 100 times. Figure 10a illustrated the overall accuracy (Level-0 and LCCS level-1 classification systems) increased for the increased percentage of training samples. It was found that the overall accuracy rapidly increased when the percentage of training samples increased from 1% to 30%, while it remained relatively stable when the percentage of training samples was higher than 30%. Therefore, the appropriate sample size should be larger than the 60,000 (30% of the total input points), fortunately, the local training samples in this study almost all exceeded the 60,000 because the training samples from neighboring 3 × 3 tiles were used to train the random forest model and classify the central tile. Similarly, Foody (2009) also found that the sample size had a positive relationship with the classification accuracy up to the point where the sample size was saturated, and Zhu et al. (2016) suggested that the optimal size was a total of 20,000 training pixels to classify an area about the size of a Landsat scene.

Secondly, many studies have demonstrated that the sample outliers had influence on the land-cover classification accuracy (Mellor et al. 2015, Pelletier et al. 2017). In this study, using previous 200,000 training points, we further analyzed the relationship between overall classification accuracy and erroneous training sample by randomly changing the category of a certain percentage of these samples and using the "noisy" samples to train the random forest classifier. Similar to the previous quantitative analysis of sample size, we gradually increased the percentage of erroneous training samples with the step of 2% and then repeat the process for 100 times. Figure 10b showed that the overall accuracy of two classification systems (level-0 and LCCS level-1) generally decreased with the increasing of percentage of erroneous sample points. It remained relatively stable when the percentage of erroneous training sample was controlled within 30%, and decreased obviously after exceeding the threshold of 30%. Meanwhile, the overall accuracy of simple classification system was more susceptible to the erroneous samples than that of the LCCS classification system in the Figure 10b. Similarly, many scientists have also demonstrated that a small number of training data have little effect on the classification results (Gong et al., 2019; Mellor et al., 2015; Pelletier et al., 2016; Zhu et al., 2016): for example, Mellor et al. (2015) found the error rate of the RF classifier was insensitive to mislabeled training data, and the overall accuracy decreased from 78.3% to 70.1% when the proportion of mislabeled training data increased from 0% to 25%. Similarly, Pelletier et al. (2016) found the RF classifier was little affected by low random noise levels up to 25%–30% but that the performance dropped at higher noise levels.

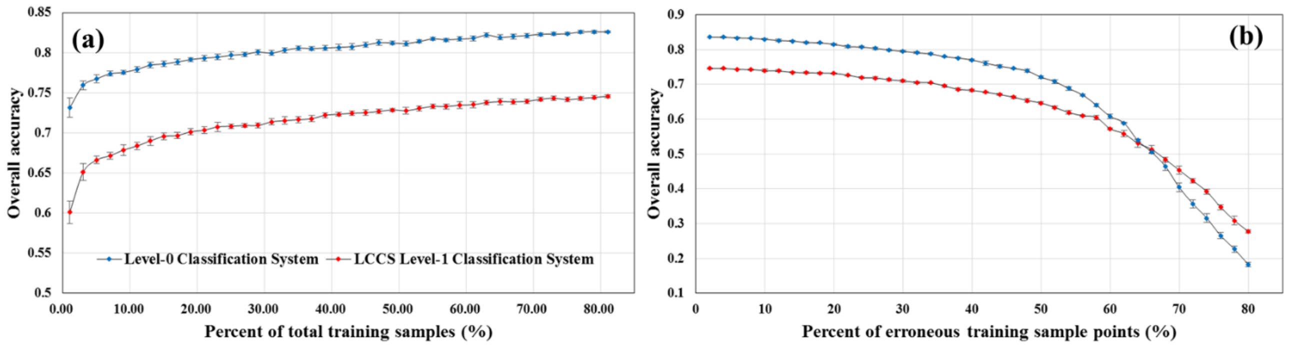

Figure 10. Sensitivity analysis showing the relations between the overall classification accuracy and the percentage of total samples and erroneous sample points.

Defourny et al. (2018) demonstrated that CCI_LC achieved an overall accuracy of 75.38% for homogeneous areas. In this study, some measures have been taken to guarantee the confidence of training samples. Some complicated land-cover types were then further optimized to improve the accuracy of the training data; for example, impervious surfaces were imported as an independent product and directly superimposed over the final global land-cover classifications, the three wetland types were merged into an overall wetland land-cover type, and four mosaicked land-cover types were removed (Table 2). After optimizing these complicated land-cover types, the overall accuracy of CCI_LC reached 77.36% for homogeneous areas based on the confusion matrix of Defourny et al. (2018). In addition, other measures, including the spectral filters applied to the MCD43A4 NBAR data, the land-cover homogeneity constraint for CCI_LC land-cover products, and the "metric centroid" algorithm for removing the resolution differences, were used to further improve confidence in the training data. Therefore, a part of training samples (exceeding 18000 points) in the previous analysis were randomly selected to quantitatively evaluate the confidence of the global training dataset, after pixel-by pixel interpretation and inspection, the validation results indicated that these samples had satisfactory performance with the overall accuracy of 91.7% for the Level-0 classification system and 82.6% for Level-1 LCCS classification system. Therefore, it can be assumed that the training data, derived by combining the MCD43A4 NBAR and CCI_LC land-cover products, were accurate and suitable for large-area land-cover mapping at 30 m.

Lastly, the sample balance is also an important factor in land-cover classification especially for rare land-cover types, because unbalanced training data would cause the under-fitting of classification model for rare land-cover types and further degrade the classification accuracy. In this study, we used the sample balancing parameters (a minimum of 600 training pixels and a maximum of 8000 training pixels per class), based on the work of Zhu et al. (2016), to alleviate the problem of unbalancing training data when deriving training samples from the GSPECLib in the Section 3.1, therefore, Figure 8 II and III illustrated that the water body, which was the rare land-cover type in the whole regions, have been accurately captured in the corresponding enlargement figures.

2) As for the relationship between land-cover heterogeneity and the classification accuracy

Except for the training sample uncertainties (including sample size, outliers) in the section 5.1, the land-cover heterogeneity also had a significant effect on the classification accuracy (Calderón-Loor et al., 2021; Wang and Liu, 2014). To clarify the relationship between land-cover heterogeneity and overall accuracy of the GLC_FCS30-2015 land-cover map, we firstly used the Shannon entropy to calculate the spatial heterogeneity using the GLC_FCS30_2015 at spatial resolution of 0.05°×0.05° (Eq. 4). Figure 11a illustrated the land-cover heterogeneity of GLC_FCS30 land-cover map. Intuitively, the highly heterogeneous regions mainly corresponded to the climatic transition zone especially for the sparse vegetation areas. Then, we combined the land-cover heterogeneity and global validation datasets (in the Section 2.3) to calculate the mean accuracy at different heterogeneity illustrated in Figure 11b. It could be found that the classification accuracy had negative relationship with land-cover heterogeneity with the slope of -0.3347, namely, the GLC_FCS30 had better performance in the homogeneous areas than that of the heterogeneous areas. Similarly, Defourny et al. (2018) also demonstrated that the CCI_LC land-cover products achieved the higher accuracy of 77.36% in the homogeneous areas than that of 75.38% in the all areas.

$$H = -\sum_{i=1}^{n}(P_i \times log_2 P_i) \qquad (4)$$

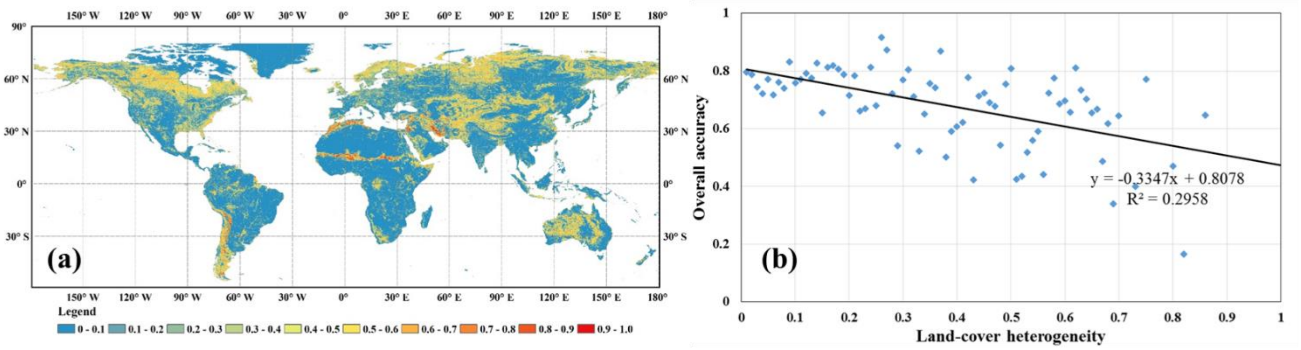

Figure 11. The land-cover heterogeneity of GLC_FCS30 land-cover map at a spatial resolution of 0.05°, and the relationship between land-cover heterogeneity and overall accuracy using the global validation datasets.

Detailed comments

Line 10: add ’a’ before ’lack’. L15: Include full names with the acronyms when they are first introduced.

Great thanks for the comment. The missed words were added throughout. The full names of the acronyms (CCI_LC and MCD43A4 NBAR) in L15 have been added.

L99: The “lack of global satellite data coverage” is no longer a challenge for MODIS and Landsat that have been free of charge for over a decade. In fact, we are now in a data-rich era, which is why supercomputing and effective data mining are critical.

Great thanks for the comment. Yes, with the free access of MODIS and Landsat imagery, the “lack of global satellite data coverage” is no longer a challenge. Therefore, the sentence has been revised as:

“Secondly, **the high cost of collecting satellite data with consistent global coverage**, the lack of the high-performance computing requirements and the difficulties in preparing image mosaics also cause problems.”

L145-146: Why not directly using the ASTER GDEM product? The most recent version 3 of GDEM has better accuracy than SRTM.

Great thanks for the useful suggestion. The GLC_FCS30-2015 land-cover maps began production in 2019, when GDEM version 3 was not integrated on the GEE platform. Based on your suggestion, our further work would use the GDEM version 3 to replace the SRTM dataset.

Section 2.3: What are your criteria for deriving how many points for each land cover class?

Great thanks for the important comment. The sample size of each land-cover type is determined by the stratified random sampling. The works of Foody et al. (2009) and Olofsson et al. (2014) have detailedly explained how to use the area proportion to calculate the appropriate validation sample size. The part has been added as:

To guarantee the confidence of the validation points, several existing prior datasets (see Table 1), high-resolution Google Earth imagery and time-series of NDVI values for each vegetated point were integrated to derive the global validation datasets. **Many studies have demonstrated that inappropriately sized validation sample could lead to limited and sometimes erroneous assessments of accuracy (Foody et al. 2009 and Olofsson et al. 2014), therefore, a stratified random sampling based on the proportion of the land-cover areas was adapted to determine the sample size of each land-cover type:**

$$\boldsymbol{n_i = n \times \frac{W_i \times p_i(1-p_i)}{\sum W_i \times p_i(1-p_i)}; \qquad n = \frac{(\sum W_i \times \sqrt{S_i(1-S_i)}\,)^2}{[S(\hat{O})]^2 + \sum W_i \times S_i(1-S_i)/N} \approx \left(\frac{\sum W_i S_i}{S(\hat{O})}\right)^2} \qquad \textbf{(1)}$$

where W_i was the area proportion for class i over the globe, S_i is the standard deviation of class i, $S(\widehat{O})$ is the standard error of the estimated overall accuracy, p_i is the expected accuracy of class i and n_i represents the sample size of the class i.

L170: Where did you get the high-resolution imagery? How many points did you check? Following what criteria?

Great thanks for the comment. The high-resolution imagery came from the Google earth software. There are 22,823 cropland validation samples in the reference dataset have been checked. Lastly, to guarantee the confidence of validation samples, all validation samples were rechecked by three experts using Google Earth software, if the rechecking results of three experts were in disagreement, the cropland point would be discarded. It has been revised as:

There are 22,823 cropland validation samples in the reference dataset (Xiong et al., 2017). In addition, due to the possible temporal interval between the acquisition of the reference data and the GLC_FCS30 products (2015), **the reference samples were checked by three interpreters using the high-resolution imagery for 2015 in the Google Earth software, and were discarded if the judgements of three experts were in disagreement. After discarding wrong cropland points and resampling using the formula (1), a total of 6,917 cropland samples in 2015 were retained.**

L177: great -> big.

Great thanks for the comment. It has been corrected.

Section: 3.1: There are multiple steps. I suggest a flowchart to describe your process. Also, how many samples did you collect for the study and for each class? What were your criteria?

Great thanks for the comment. According to your suggestion, the flowchart has been added, and the sample sizes of each land-cover type are calculated by the area proportion. Specifically, the part has been supplemented as:

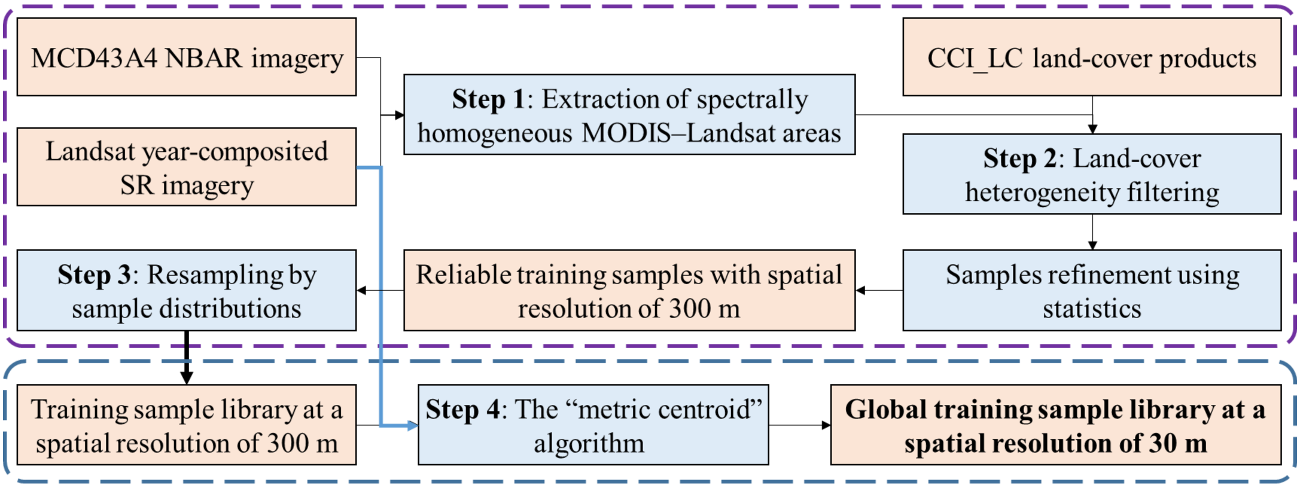
MCD43A4 NBAR imagery
Landsat year-composited SR imagery
Step 1: Extraction of spectrally homogeneous MODIS–Landsat areas
CCI_LC land-cover products
Step 2: Land-cover heterogeneity filtering
Samples refinement using statistics
Reliable training samples with spatial resolution of 300 m
Step 3: Resampling by sample distributions
Training sample library at a spatial resolution of 300 m
Step 4: The "metric centroid" algorithm
Global training sample library at a spatial resolution of 30 m

Figure 3. The flowchart of deriving training samples by using multi-source datasets.

Similar to our previous works (Zhang et al., 2019; Zhang et al., 2018), four key steps were adopted to guarantee the confidence of each training point, as illustrated in the Figure 3. As in Zhang et al. (2019), the spectrally homogeneous MODIS–Landsat areas were firstly identified based on the variance of a 3×3 local window using spectral thresholds of [0.03, 0.03, 0.03, 0.06, 0.03, and 0.03] for the six MODIS bands (blue, green, red, NIR, SWIR1, and SWIR2) in the both MCD43A4 NBAR products and Landsat SR imagery (Feng et al., 2012). It should be noted that the year-composited Landsat SR data were downloaded from GEE platform with the sinusoidal projection. As the MCD43A4 NBAR is corrected for view-angle effects and Landsat has a small view angle of ±7.5°, the view-angle difference between MCD43A4 and Landsat SR could be considered negligible.

Before the process of refinement and labeling, the CCI_LC land-cover products, which had geographical projections, were reprojected to the sinusoidal projection of MCD43A4. The spatial resolution of MCD43A4 was 1.67 times that of the CCI_LC land-cover product and the spectrally homogeneous MODIS–Landsat areas had been identified in the 3×3 local windows. Also, Defourny et al. (2018) and Yang et al. (2017b) found that the CCI_LC performed better over homogeneous areas; therefore, a larger local 5×5 window was applied to the CCI_LC land-cover product to refine and label each spectrally homogeneous MODIS-Landsat pixel. Specifically, the land-cover heterogeneity in the local 5×5 window was calculated as being the percentages of land-cover types occurring within the window (Jokar Arsanjani et al., 2016a). Aware of the possibility of reprojection and classification errors in the CCI_LC products, the land-cover heterogeneity threshold was empirically selected as approximately 0.95; in other words, if the maximum frequency of dominant land-cover types was less than 22 in the 5×5 window, the point was excluded from GSPECLib. After a spatial–spectral filter had been applied to MCD43A4 and a heterogeneity filter to the CCI_LC product, the points that had homogeneous spectra and land-cover types were retained. In addition, to further remove the abnormal points contaminating by classification error in the CCI_LC, the homogeneous points were refined based on their spectral statistics distribution, in which the normal samples would form the peak of the distribution whereas the influenced samples were on the long tail (Zhang et al., 2018). It should be noted that the geographical coordinates of each homogeneous point were selected as being the center of the local window in the CCI_LC product because this had a higher spatial resolution than that of MCD43A4.

Then, Zhu et al. (2016) and Jin et al. (2014) found that the distribution (proportional to area and equal allocation) and balance of training data had significant impact on classification results, and quantitatively demonstrated that the proportional approach usually achieve higher overall accuracy than the equal allocation distribution. In addition, Zhu et al. (2016) also suggested to extract a minimum of 600 training pixels and a maximum of 8000 training pixels per class for alleviating the problem of unbalancing training data. In this study, the proportional distribution and sample balancing parameters were used to resample these homogeneous points in each GSPECLib 158.85 km×158.85 km geographic grid cell.

L214: land-covers -> land cover.

Great thanks for the comment. It has been corrected.

L303-304: I do not agree that "classification accuracy was insensitive to these parameters". Please see a review of RF in RS classification by Belgiu and Dragut (2016).

Great thanks for the comment. The statement has been revised based on the work of Belgiu and Dragut (2016) as:

“Belgiu et al. (2016) also explained that the classification accuracy was less sensitive to Ntree than to the Mtry parameter, and Mtry was usually set to the square root of the number of input variables. Due to these advantages, the RF classifier is widely used in land-cover mapping”

L320-322: It is vague how you balanced performance, efficiency, and sample volumes. What criteria did you use?

Great thanks for the comment. The reason why we choose the 5°×5° geographical tiles as the mapping unit is because our experiments and the works of Zhang et al. (2017) found that if we chose the 170 km×180 km (the Landsat size) as a spatial unit, there will be lacking of training samples for sparse land-cover types. A good solution is to import some training samples from neighboring 3 by 3 tiles if the training samples are insuffient (Zhang and Roy, 2017; Zhang et al., 2019). Therefore, the 5°×5° geographical tiles, approximately 3×3 Landsat scenes, to avoid the under-fitting when training the local adaptive model. 2) As the GEE has some limitations for computation capability and memory, if we choose bigger spatial unit, the GEE platform would have some over-memory/over-time errors. The sentences have been added as:

“Furthermore, **as illustrated in the previous works, the training samples in a small spatial grid (Landsat scene) were not enough especially for sparse land-cover types, and the training samples from neighboring 3 by 3 tiles were also imported (Zhang and Roy, 2017; Zhang et al., 2019), as well as GEE platform had some limitations for computation capacity and memory**. Therefore, after balancing the accuracy performance, computation efficiency and training sample volume, the local adaptive random forest models, which split the globe into approximately 948 5°×5° geographical tiles (approximately 3×3 Landsat scenes) similar to our previous work (Zhang et al., 2020), were applied to generate a lot of regional land-cover maps."

Zhang, H. K. and Roy, D. P.: Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification, Remote Sensing of Environment, 197, 15-34, https://doi.org/10.1016/j.rse.2017.05.024, 2017.
Zhang, X., Liu, L., Chen, X., Xie, S., and Gao, Y.: Fine Land-Cover Mapping in China Using Landsat Datacube and an Operational SPECLib-Based Approach, Remote Sensing, 11, 1056, https://doi.org/10.3390/rs11091056, 2019.

Section 5.1: "huge training samples". Exactly how many samples were used? It is vague to use "exceeded 20 million points".

Great thanks for the comment. The exact samples of **27,858,258 points** has been added as:

"In addition, in contrast to other studies that used manual interpretation of samples for global land-cover mapping (Friedl et al., 2010; Gong et al., 2013; Tateishi et al., 2014), the total number of training samples in this study **reaching 27,858,258 points** and so was tens to hundreds of times higher than that used in these global land-cover classifications."

Since building the training sample database is the most important contribution of the project, it is critical and would certainly benefit the users through discussing how the number of the training samples and how sample balance (across classes) have affected the results. The authors lightly touched on the outlier effect, but there is a lack of in-depth analysis and discussion using the data from the present project.

Great thanks for the comment. The effects of training sample sizes and outlier effect have been added in the manuscript in the Discussion Section as:

To demonstrate the importance of sample sizes, 200,000 points, approximately 1% of total training samples, were randomly selected to quantitatively analyse the relationship between overall accuracy and the corresponding sample size. Specifically, we used the 10-fold cross-validation method to split these points into training and validation samples, and then gradually increase the size of training samples with the step of 2% and repeat the process for 100 times. Figure 10a illustrated the overall accuracy (Level-0 and LCCS level-1 classification systems) increased for the increased percentage of training samples. It was found that the overall accuracy rapidly increased when the percentage of training samples increased from 1% to 30%, while it remained relatively stable when the percentage of training samples was higher than 30%. Therefore, the appropriate sample size should be larger than the 60,000 (30% of the total input points), fortunately, the local training samples in this study almost all exceeded the 60,000 because the training samples from neighboring 3 × 3 tiles were used to train the random forest model and classify the central tile. Similarly, Foody (2009) also found that the sample size had a positive relationship with the classification accuracy up to the point where the sample size was saturated, and Zhu et al. (2016) suggested that the optimal size was a total of 20,000 training pixels to classify an area about the size of a Landsat scene.

Secondly, many studies have demonstrated that the sample outliers had influence on the land-cover classification accuracy (Mellor et al. 2015, Pelletier et al. 2017). In this study, using previous 200,000 training points, we further analyzed the relationship between overall classification accuracy and erroneous training sample by randomly changing the category of a certain percentage of these samples and using the “noisy” samples to train the random forest classifier. Similar to the previous quantitative analysis of sample size, we gradually increased the percentage of erroneous training samples with the step of 2% and then repeat the process for 100 times. Figure 10b showed that the overall accuracy of two classification systems (level-0 and LCCS level-1) generally decreased with the increasing of percentage of erroneous sample points. It remained relatively stable when the percentage of erroneous training sample was controlled within 30%, and decreased obviously after exceeding the threshold of 30%. Meanwhile, the overall accuracy of simple classification system was more susceptible to the erroneous samples than that of the LCCS classification system in the Figure 10b. Similarly, many scientists have also demonstrated that a small number of erroneous training data have little effect on the classification results (Gong et al., 2019; Mellor et al., 2015; Pelletier et al., 2016; Zhu et al., 2016): for example, Mellor et al. (2015) found the error rate of the RF classifier was insensitive to mislabeled training data, and the overall accuracy decreased from 78.3% to 70.1% when the proportion of mislabeled training data increased from 0% to 25%. Similarly, Pelletier et al. (2016) found the RF classifier was little affected by low random noise levels up to 25%–30% but that the performance dropped at higher noise levels.

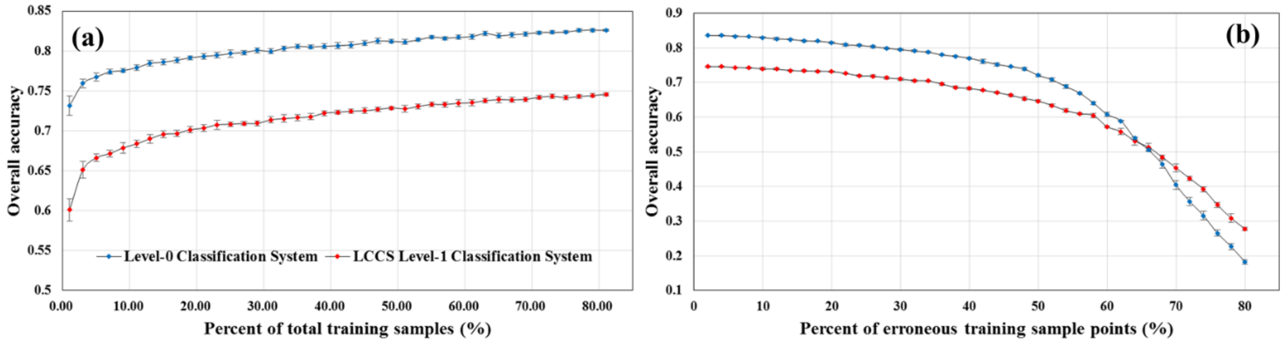

Figure 10. Sensitivity analysis showing the relations between the overall classification accuracy and the percentage of total samples and erroneous sample points.

Defourny et al. (2018) demonstrated that CCI_LC achieved an overall accuracy of 75.38% for homogeneous areas. In this study, some measures have been taken to guarantee the confidence of training samples. Some complicated land-cover types were then further optimized to improve the accuracy of the training data; for example, impervious surfaces were imported as an independent product and directly superimposed over the final global land-cover classifications, the three wetland types were merged into an overall wetland land-cover type, and four mosaicked land-cover types were removed (Table 2). After optimizing these complicated land-cover types, the overall accuracy of CCI_LC reached 77.36% for homogeneous areas based on the confusion matrix of Defourny et al. (2018). In addition, other measures, including the spectral filters applied to the MCD43A4 NBAR data, the land-cover homogeneity constraint for CCI_LC land-cover products, and the “metric centroid” algorithm for removing the resolution differences, were used to further improve confidence in the training data. Therefore, a part of training samples (exceeding 18000 points) in the previous analysis were randomly selected to quantitatively evaluate the confidence of the global training dataset, after pixel-by pixel interpretation and inspection, the validation results indicated that these samples had satisfactory performance with the overall accuracy of 91.7% for the Level-0 classification system and 82.6% for Level-1 LCCS classification system. Therefore, it can be assumed that the training data, derived by combining the MCD43A4 NBAR and CCI_LC land-cover products, were accurate and suitable for large-area land-cover mapping at 30 m.

As for the issue of sample balance (across classes), our training samples have considered the factor in the Section 3.1 as:

Then, Zhu et al. (2016) and Jin et al. (2014) found that the distribution (proportional to area and equal allocation) and balance of training data had significant impact on classification results, and quantitatively demonstrated that the proportional approach usually achieve higher overall accuracy than the equal allocation distribution. In addition, Zhu et al. (2016) also suggested to extract a minimum of 600 training pixels and a maximum of 8000 training pixels per class for alleviating the problem of unbalancing training data. In this study, the proportional distribution and sample balancing parameters were used to resample these homogeneous points in each GSPECLib 158.85 km×158.85 km geographic grid cell.

The issue of sample balance has also been discussed in the Discussion as:

Lastly, the sample balance is also an important factor in land-cover classification especially for rare land-cover types, because unbalanced training data would cause the under-fitting of classification model for rare land-cover types and further degrade the classification accuracy. In this study, we used the sample balancing parameters (a minimum of 600 training pixels and a maximum of 8000 training pixels per class), based on the work of Zhu et al. (2016), to alleviate the problem of unbalancing training data when deriving training samples from the GSPECLib in the Section 3.1, therefore, Figure 8 II and III illustrated that the water body, which was the rare land-cover type in the whole regions, have been accurately captured in the corresponding enlargement figures.